

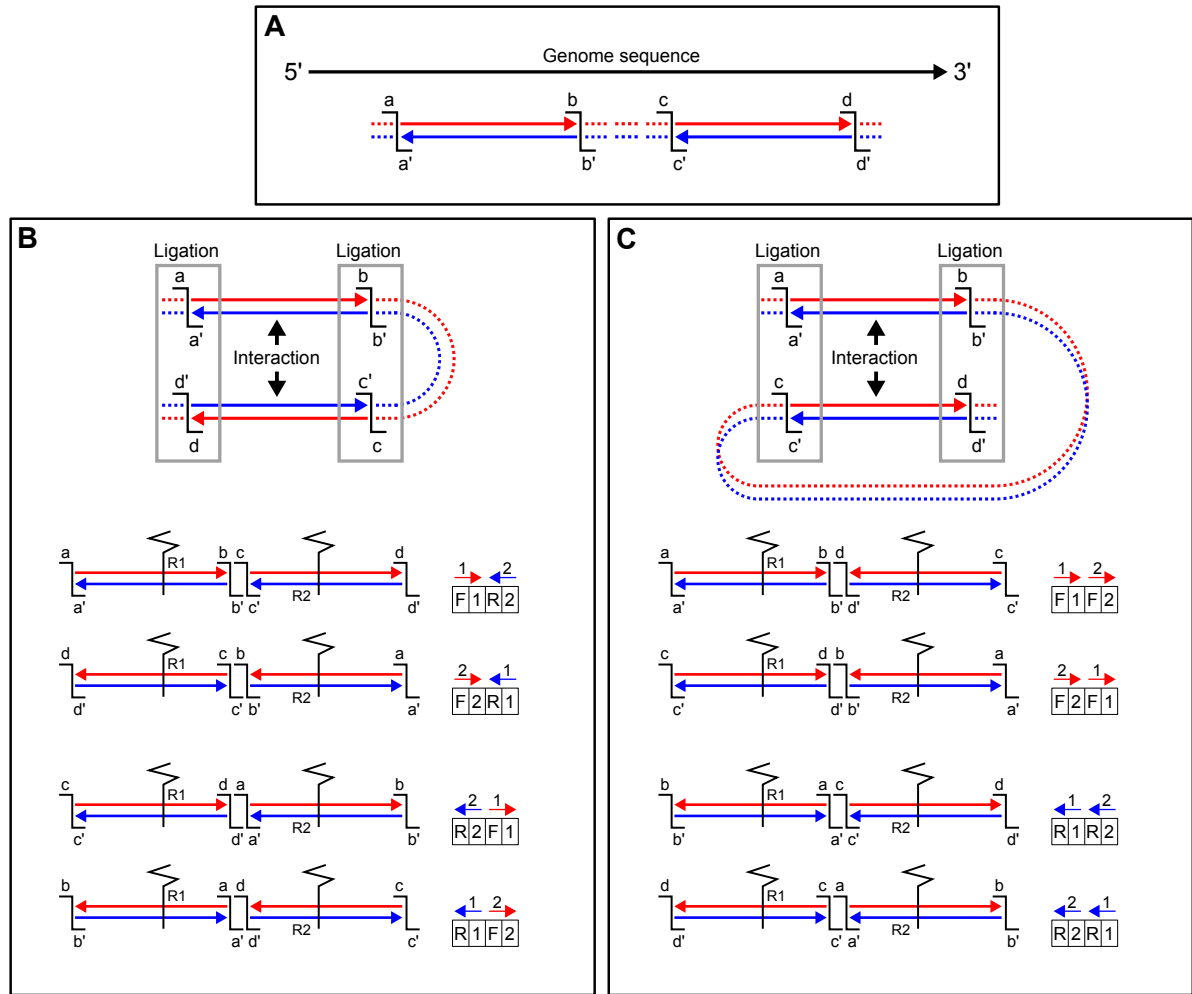
# Using paired-end read orientations to assess and mitigate technical biases in capture Hi-C

## - Supplementary Figures -

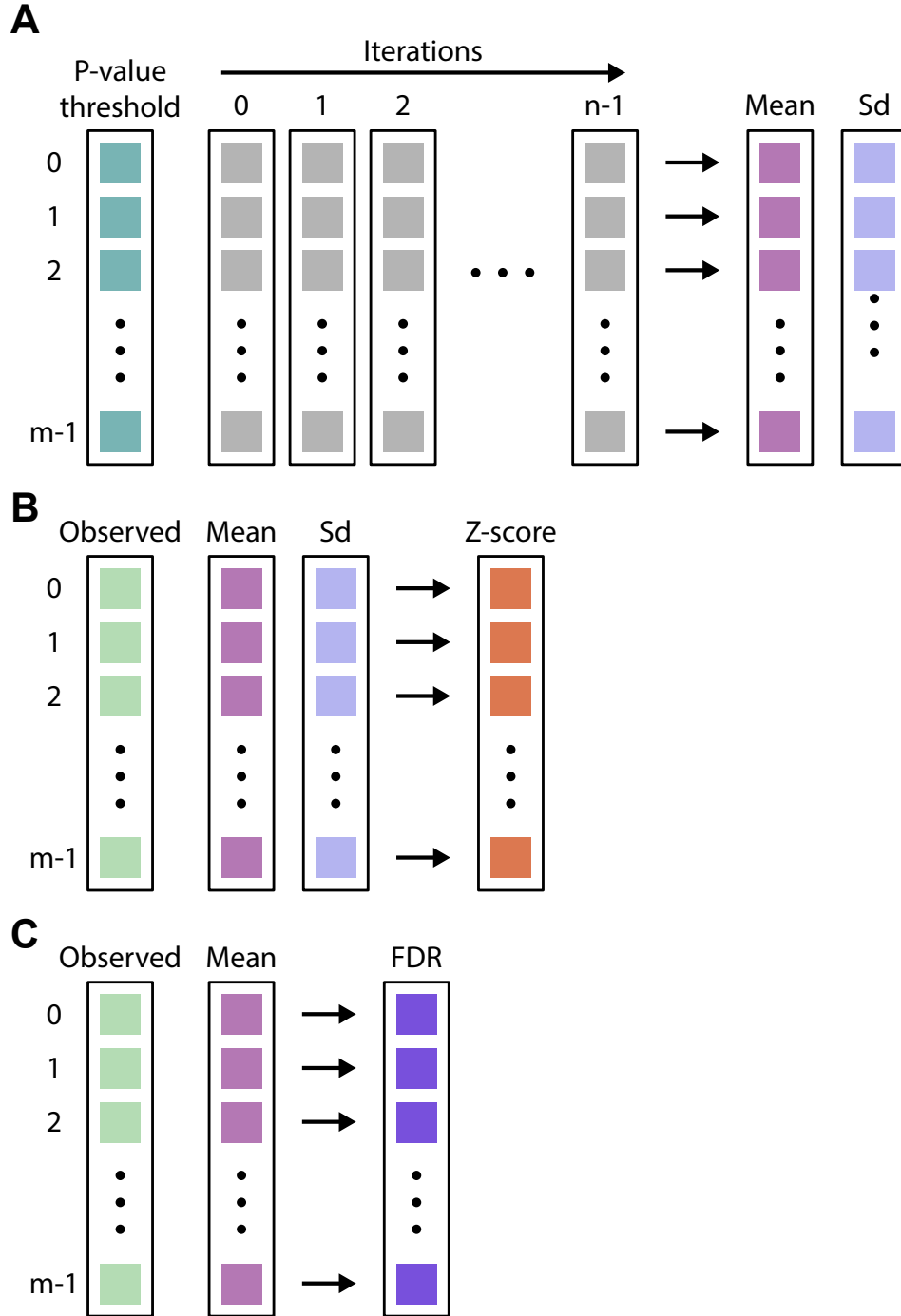
Peter Hansen<sup>1,2,\*</sup>, Hannah Blau<sup>1</sup>, Jochen Hecht<sup>3</sup>, Guy Karlebach<sup>1</sup>,  
Alexander Krannich<sup>4</sup>, Robin Steinhaus<sup>5,6</sup>, Matthias Truss<sup>7</sup>, and Peter N. Robinson<sup>1,2</sup>

## Contents

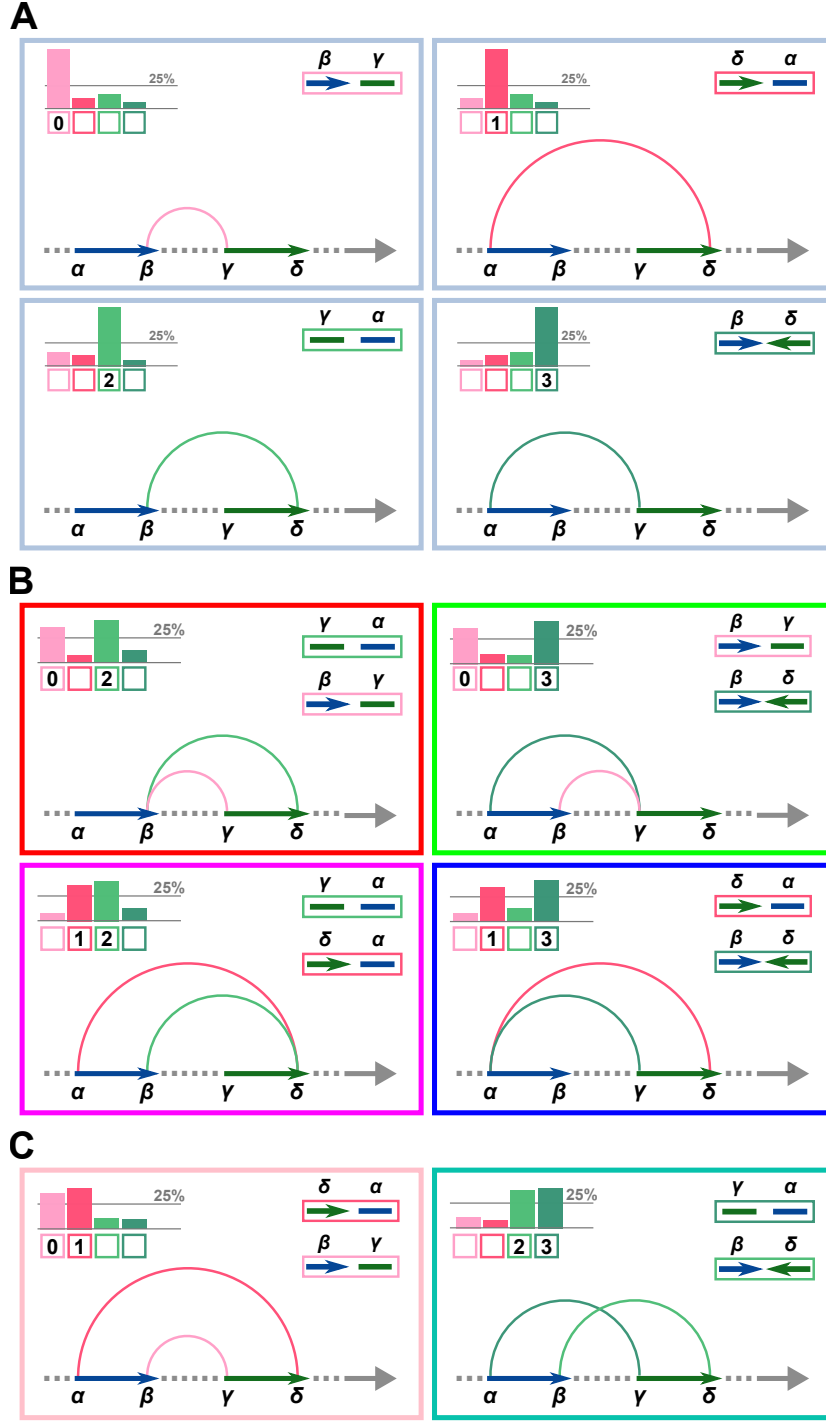
<b>Figure S1</b> Re-ligations, chimeric fragments, and mapped paired-end read orientations . . . . .	2
<b>Figure S2</b> Selection of unbalanced interactions at a chosen FDR threshold . . . . .	3
<b>Figure S3</b> Configurations of unbalanced interactions . . . . .	4
<b>Figure S4</b> Enrichment of regulatory elements within other-ends - Procedure validation . . . . .	5
<b>Figure S5</b> Overlaps of BFC0, BFC1, and BFC2 fragments . . . . .	6
<b>Figure S6</b> Decreasing efficiency of baits with increasing distance from the restriction site . . . . .	7
<b>Figure S7</b> Repeat content of baits for hematopoietic cell dataset . . . . .	8
<b>Figure S8</b> Unbaited fragment analysis for the pooled Hi-C dataset . . . . .	9
<b>Figure S9</b> Impact of technical biases reflected in count imbalances - CHi-C - MAC-M0 - EN . . . . .	10
<b>Figure S10</b> Impact of technical biases reflected in count imbalances - Pooled Hi-C dataset . . . . .	11
<b>Figure S11</b> Enrichment of regulatory elements within other-ends - Without bait-to-bait . . . . .	12
<b>Figure S12</b> Unbalanced interactions and their configurations - mESCs . . . . .	13
<b>Figure S13</b> Overlaps of BFC0, BFC1, and BFC2 fragments - mESCs . . . . .	14
<b>Figure S14</b> Bait analysis - mESCs . . . . .	15
<b>Figure S15</b> Impact of technical biases reflected in count imbalances - mESCs . . . . .	16
<b>Figure S16</b> Median total read pair counts and distances of interactions - mESCs . . . . .	17



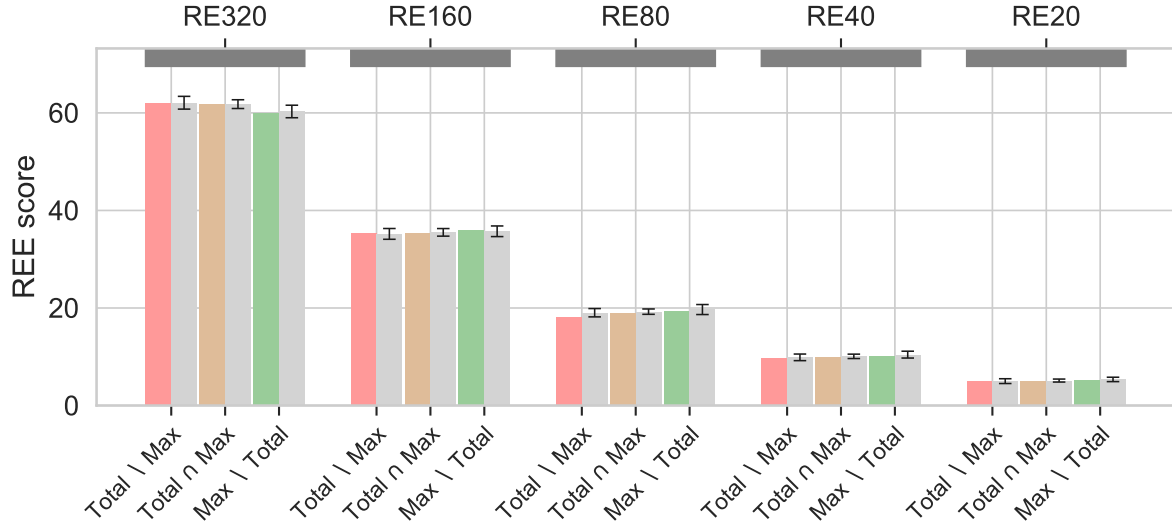
**Figure S1. Re-ligations, chimeric fragments, and mapped paired-end read orientations.** Alternative schematic representation to that shown in Figure 1 in the main text. **(A)** Two restriction fragments arranged sequentially on the genomic sequence. **(B and C)** There are four different ways in which two restriction fragments can re-ligate with each other (gray boxes). Fragmentation of DNA introduces random breakpoints (jagged symbols), resulting in chimeric fragments containing DNA from different genomic regions separated by a ligation site in the center. With paired-end sequencing, chimeric fragments are sequenced from the random breakpoints towards the center, on one side the forward strand and on the other side the reverse strand. Mapping the paired-end reads to the reference sequence results in read pairs with different relative orientations depending on which ends of the two restriction fragments have re-ligated. **(B)** If the ends of the restriction fragments facing each other in the genomic sequence ligate (b and c), this results in mapped paired-end reads that point inwards. If the ends of the restriction fragments that face away from each other in the genomic sequence ligate (a and d), this results in mapped paired-end reads that point outwards. **(C)** If either the two 5' ends (a and c) or the two 3' ends (b and d) of the restriction fragments re-ligate, this results in paired-end reads that are mapped to the same strand, either to the forward (b and d) or to the reverse strand (a and c). This is because the two DNA segments in the chimeric fragments have opposite orientations relative to the reference sequence. Therefore, only the reverse complement can be mapped for one of the two paired-end reads.



**Figure S2. Selection of unbalanced interactions at a chosen FDR threshold.** We use a randomization procedure to select unbalanced interactions at a chosen FDR threshold. In each iteration, we randomize the four read pair counts of each interaction according to a uniform distribution and then determine, at each of the  $m$  P-value thresholds from a given range, the number of interactions that are still classified as unbalanced. After all  $n$  iterations are complete, we calculate the mean and standard deviation of unbalanced interaction counts for each P-value threshold. We also determine the number of interactions that are classified as unbalanced given the original data. From these values, we calculate Z-scores and estimate FDRs (Methods). To select unbalanced interactions at a chosen FDR threshold, we use the largest P-value threshold for which the estimated FDR is below the chosen FDR threshold.

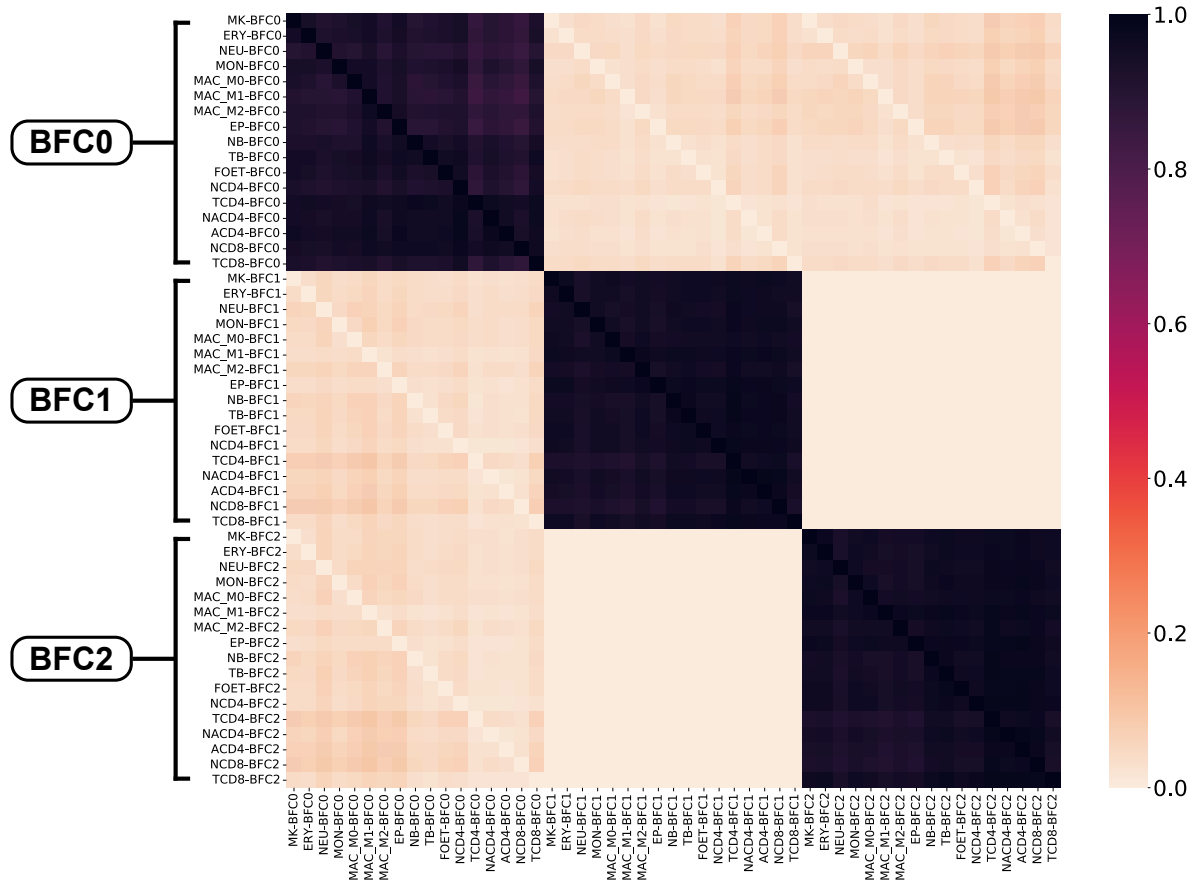


**Figure S3. Configurations of unbalanced interactions.** We distinguish 10 configurations of unbalanced interactions, depending on which of the four relative paired-end read orientations predominate. We assigned different colors to the configurations (color of the boxes). **(A)** In four cases (0X, 1X, 2X, 3X), only one paired-end read orientation predominates, with two fragment ends involved. **(B)** In four other cases (02, 03, 12, 13), two paired-end read orientations predominate, with three fragment ends involved. **(C)** And in two cases (01, 23), also two paired-end read orientations predominate, but all four fragment ends are involved.

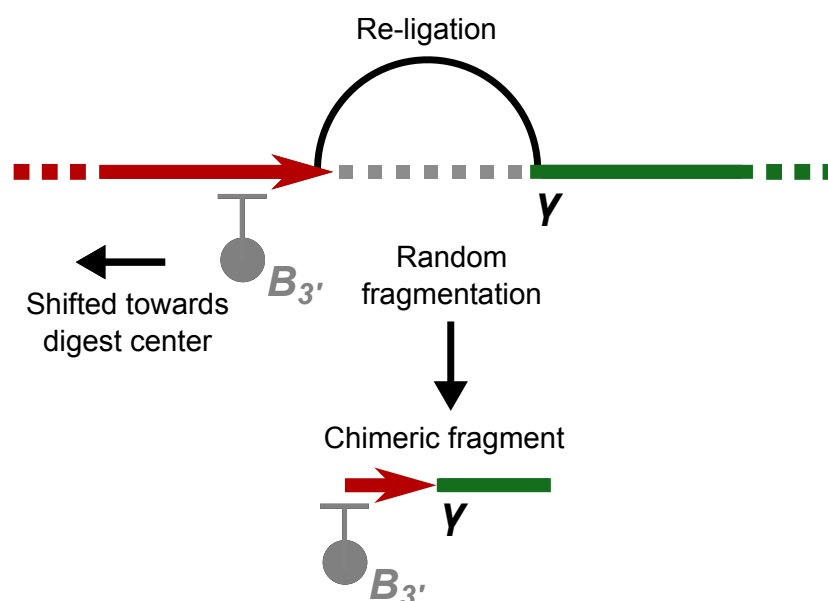


**Figure S4. Enrichment of regulatory elements within other-ends - Procedure validation.**

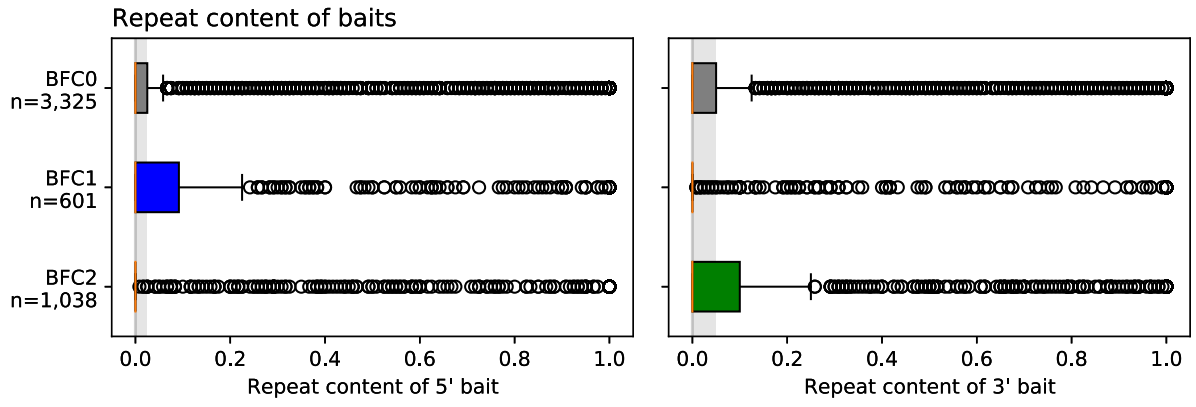
To validate our randomization procedure for assessing the enrichment of regulatory elements within other-ends of interaction (Methods), we generated data for five categories of random regulatory elements (RE320, ..., RE20). For each category, we randomly selected genomic regions with a length of 50 to 1000 bp with the categories differing in the number of selected elements (320,000, ..., 20,000). The horizontal dark grey bars at the top indicate the category to which the REE scores shown below refer. When we use these random categories instead of ENCODE's cCREs and ENA2 enhancers to analyze the interactions obtained from the MAC-M0 cell type dataset of Javierre et al. 2016 based on total counts only (light red), both total and maximum counts (beige), and maximum counts only (green) using CHiCAGO, the REE scores within categories are approximately equal, and our randomization procedure does not detect any significant enrichment compared to REE scores obtained after randomization of other-ends (light gray).



**Figure S5. Overlaps of BFC0, BFC1, and BFC2 fragments.** Based on the ratio of class 2 and class 3 paired-end reads at baited fragments, we subdivided the fragments into classes BFC0, BFC1, and BFC2 (Methods) and determined the proportion of pairwise overlap for each of the 17 hematopoietic cell types and each of the three fragment classes. A box in row  $i$  and column  $j$  indicates the proportion of fragments in set  $i$  that are also in set  $j$ .

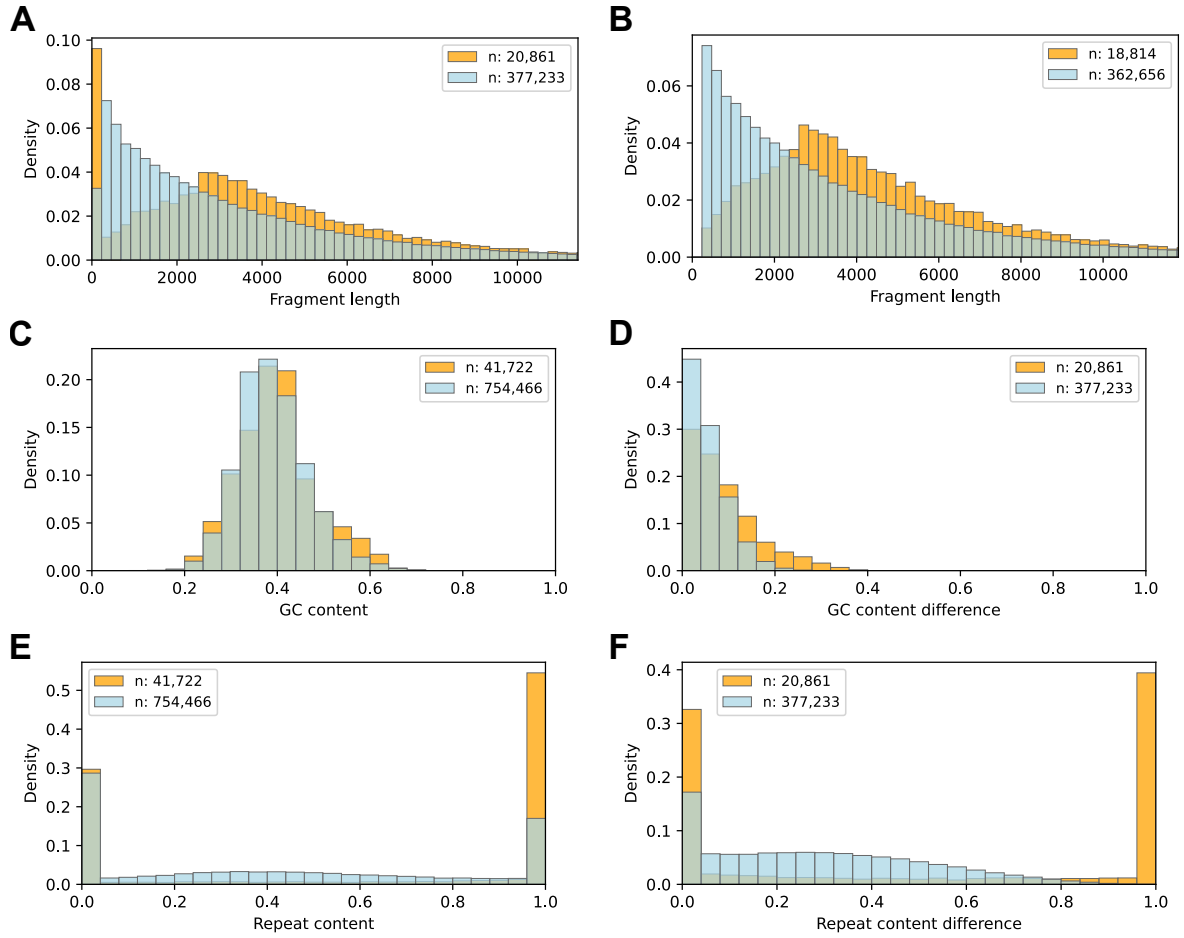


**Figure S6. Decreasing efficiency of baits with increasing distance from the restriction site.** For CHi-C, the baits are ideally placed right next to the restriction sites of fragments to be enriched, because then the target sequence is most likely to occur in chimeric fragments. However, in some cases bait selection criteria require baits to be shifted towards the center of restriction fragments, for example to reduce the GC or repeat content of the bait sequence. This in turn means that the complete target sequence of the bait is less likely to occur in chimeric fragments, making the bait less efficient or even ineffective depending on the distance between the target sequence and the restriction site.

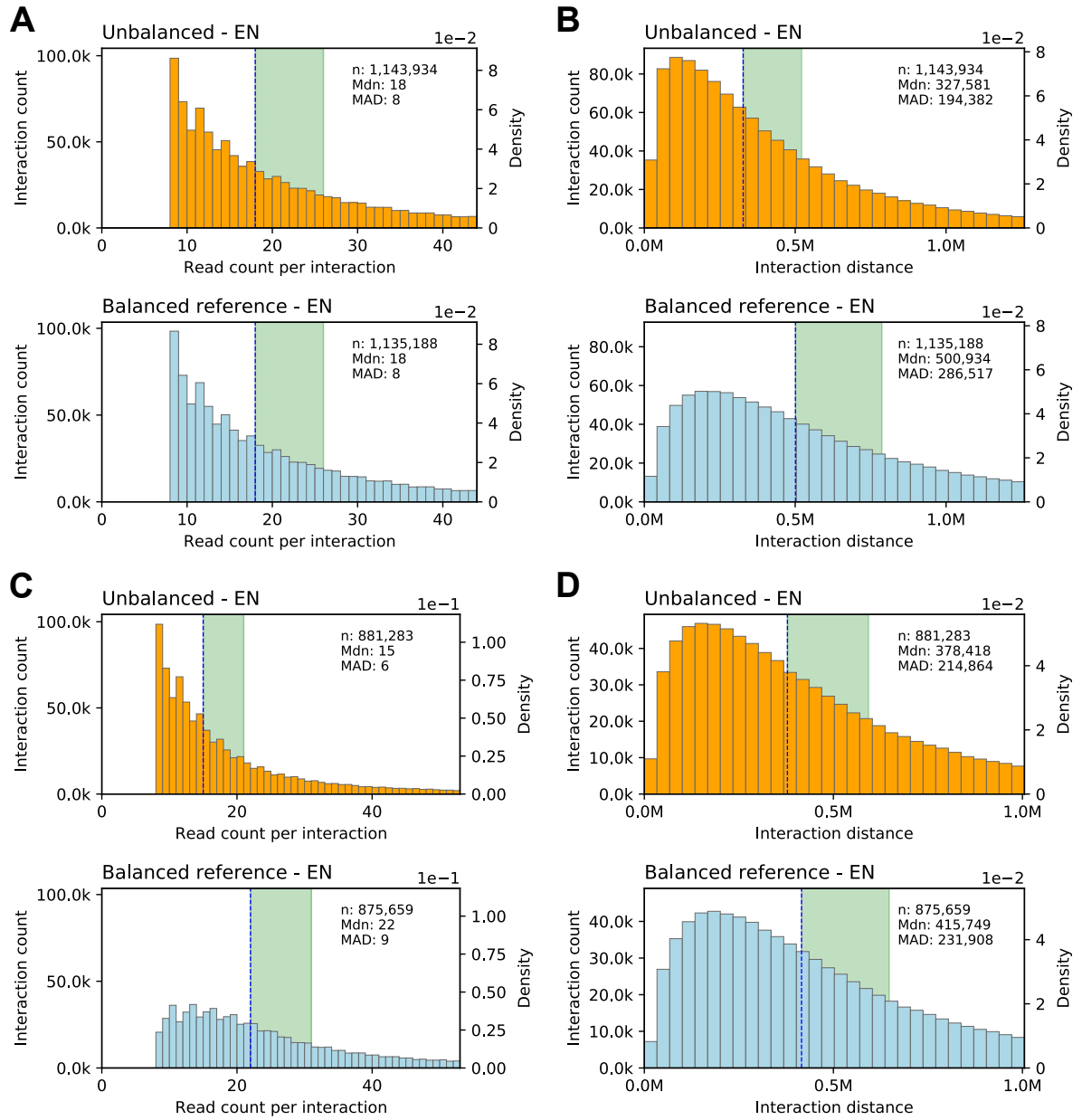


**Figure S7. Repeat content of baits for hematopoietic cell dataset.** Distributions of repeat content of baits at the 5' and 3' ends of BFC0, BFC1, and BFC2 fragments with unshifted baits at both ends. The repeat content corresponds to the proportion of bases that are masked by RepeatMasker (Methods).

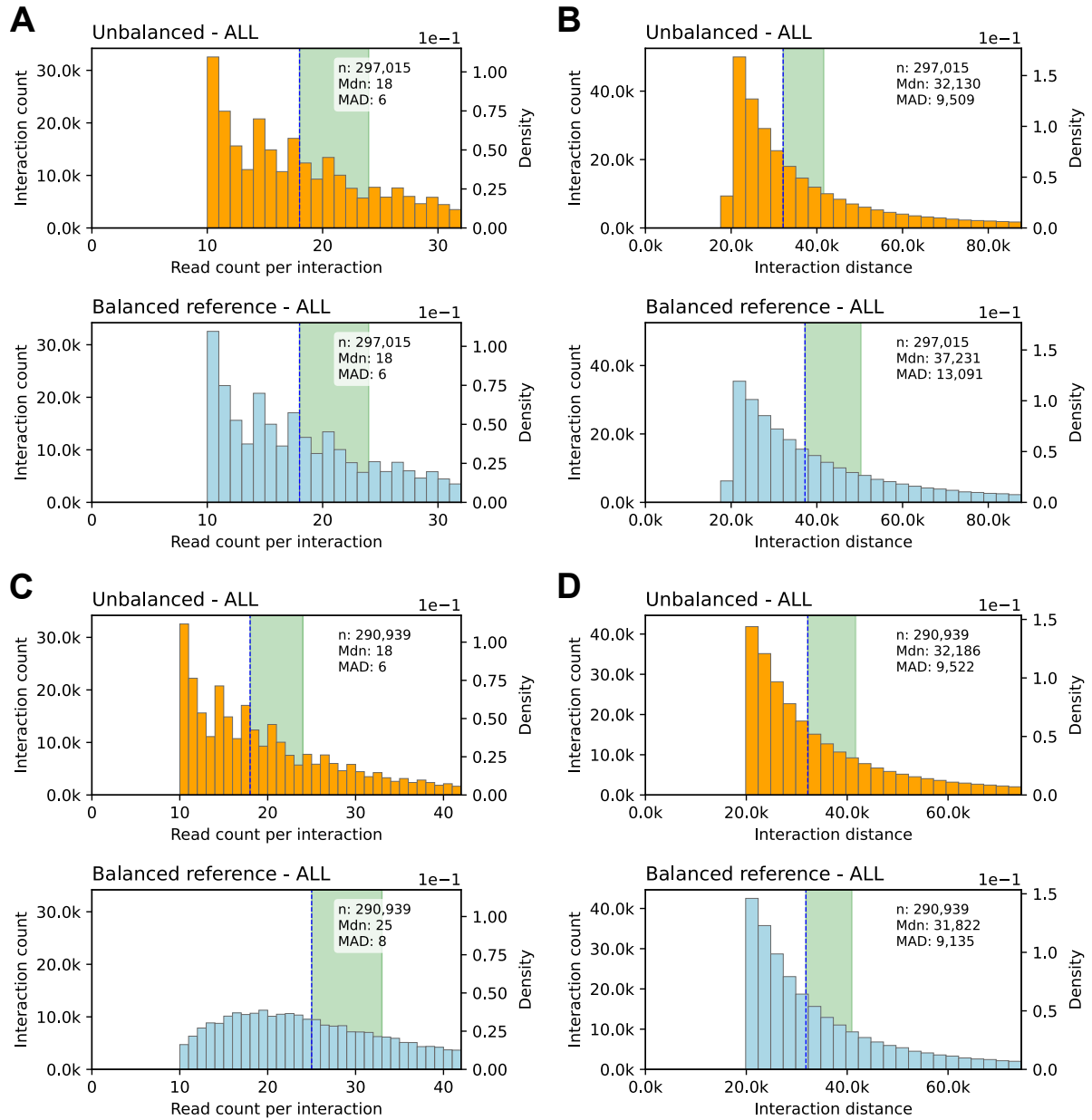




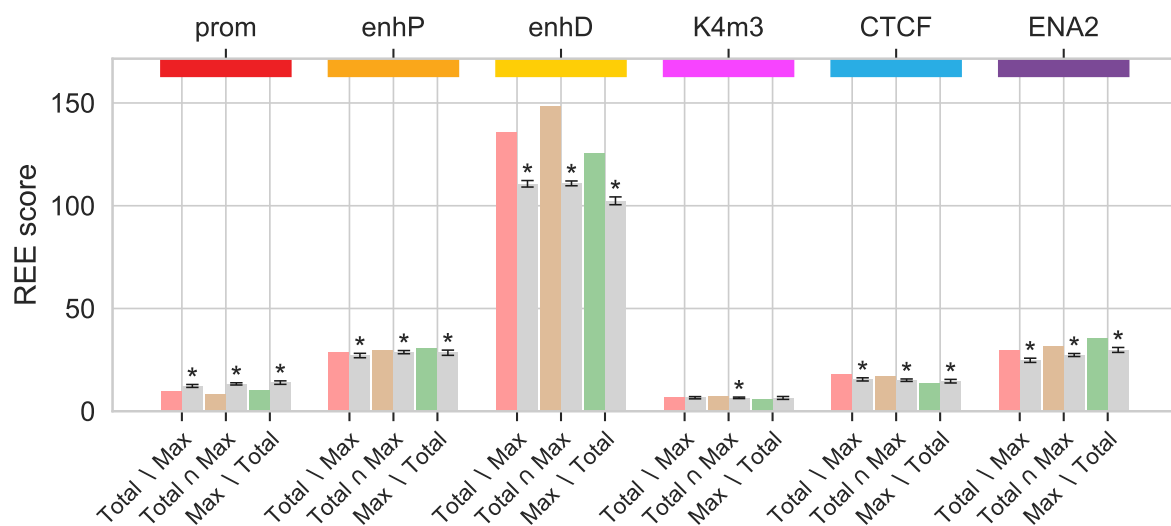
**Figure S8. Unbaited fragment analysis for the pooled Hi-C dataset.** For the Hi-C dataset, for which we pooled interactions across eight hematopoietic cell types, we performed the same analysis as for the ChIP dataset for cell type MAC-M0 (Figure 4 in the main text). However, in this case we considered all restriction fragments involved in interactions and not only the ‘other end’ fragments (N) of interactions with enrichment status NE or EN.



**Figure S9. Impact of technical biases reflected in count imbalances - CHi-C - MAC-M0 - EN.** The histograms shown are analogous to those in Figure 5 in the main text, but they were generated for EN rather than NE interactions.

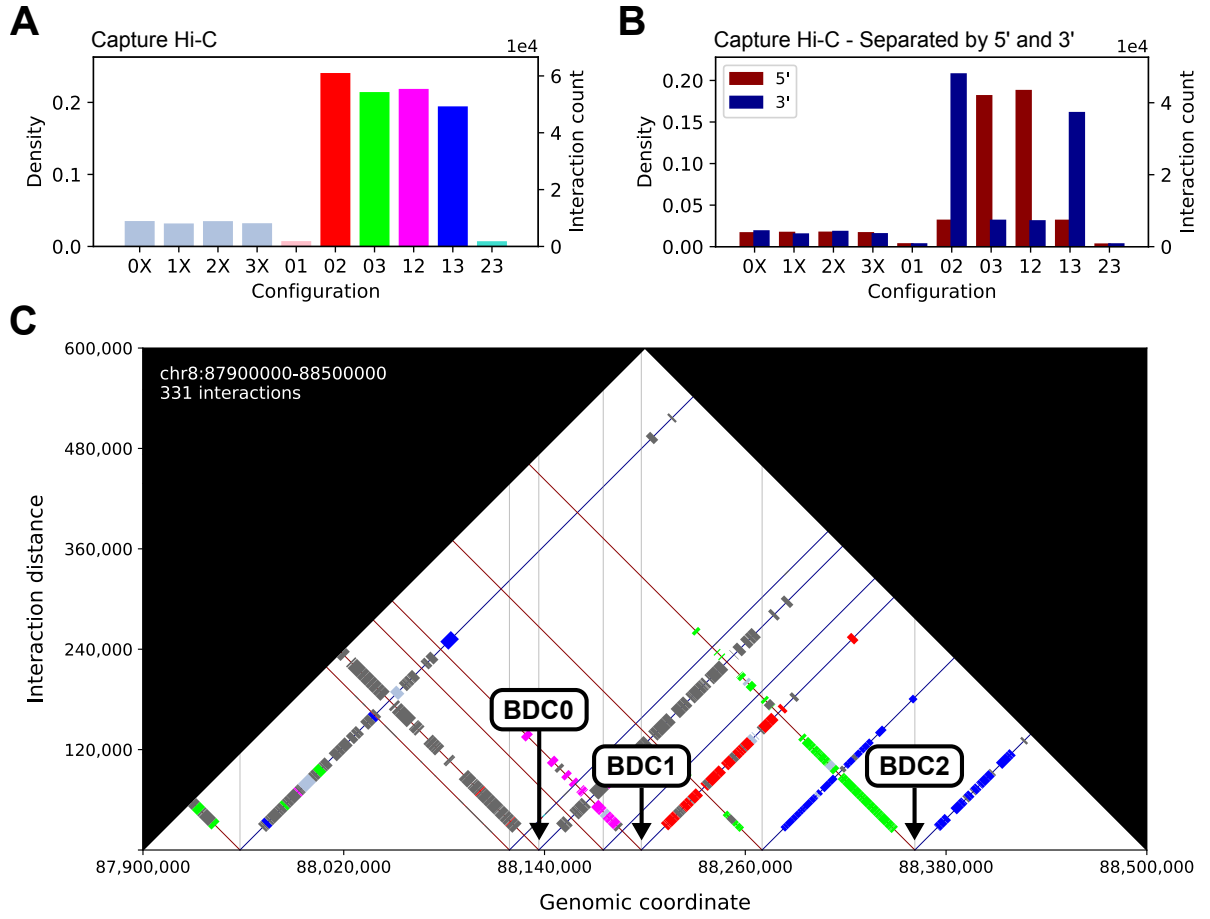


**Figure S10. Impact of technical biases reflected in count imbalances - Pooled Hi-C dataset.** We performed the analysis of the CHi-C datasets with the distance-dependent contact frequencies (Figure 5 and 6 in the main text) also for the Hi-C dataset, for which we pooled interactions across eight hematopoietic cell types. However, in the case of Hi-C, we performed the analysis regardless of enrichment status, since all interactions have enrichment status NN. For CHi-C, we only included interactions with enrichment status NE or EN in the analysis, which together form the vast majority of CHi-C interactions.

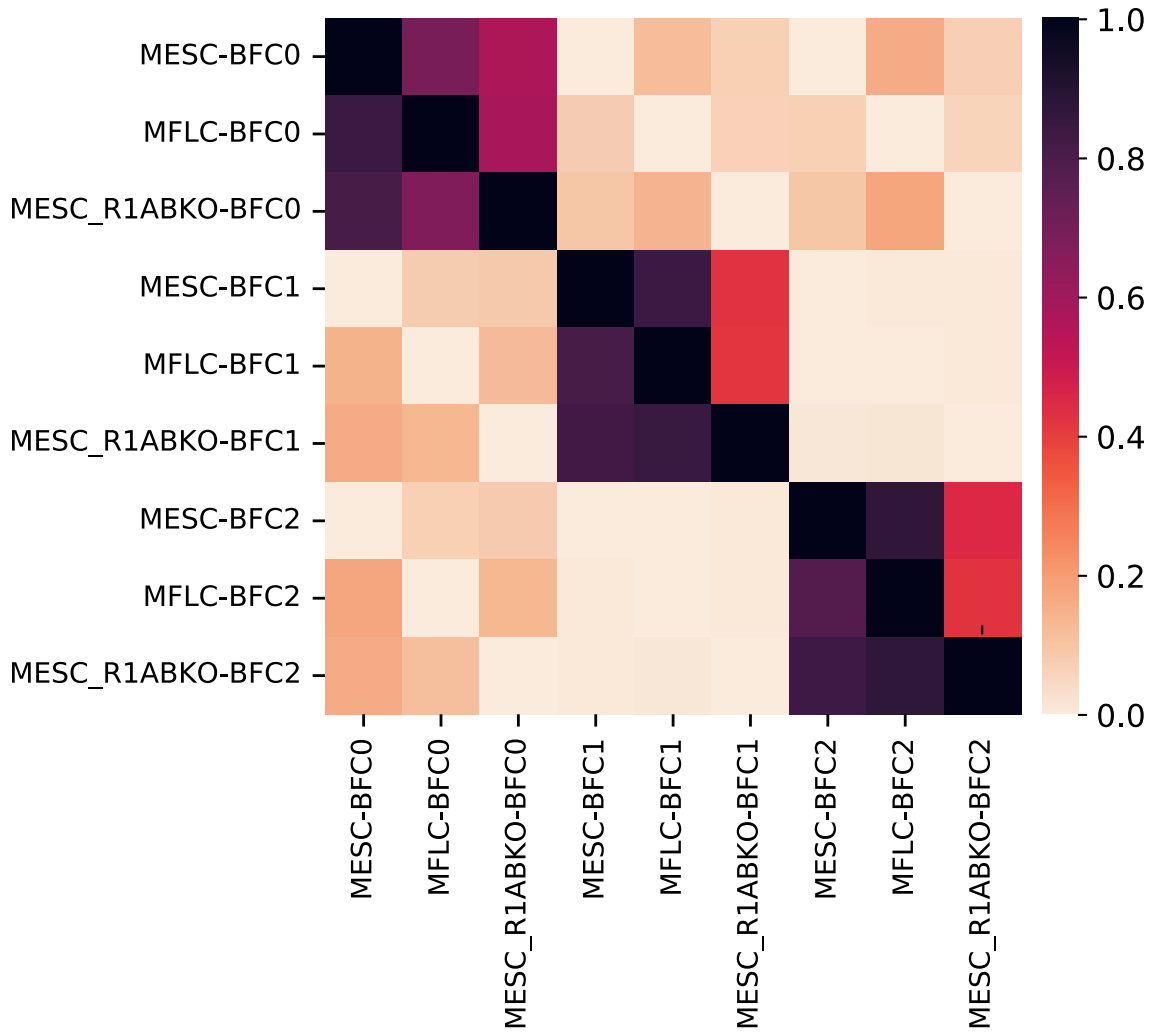


**Figure S11. Enrichment of regulatory elements within other-ends - Without bait-to-bait.**

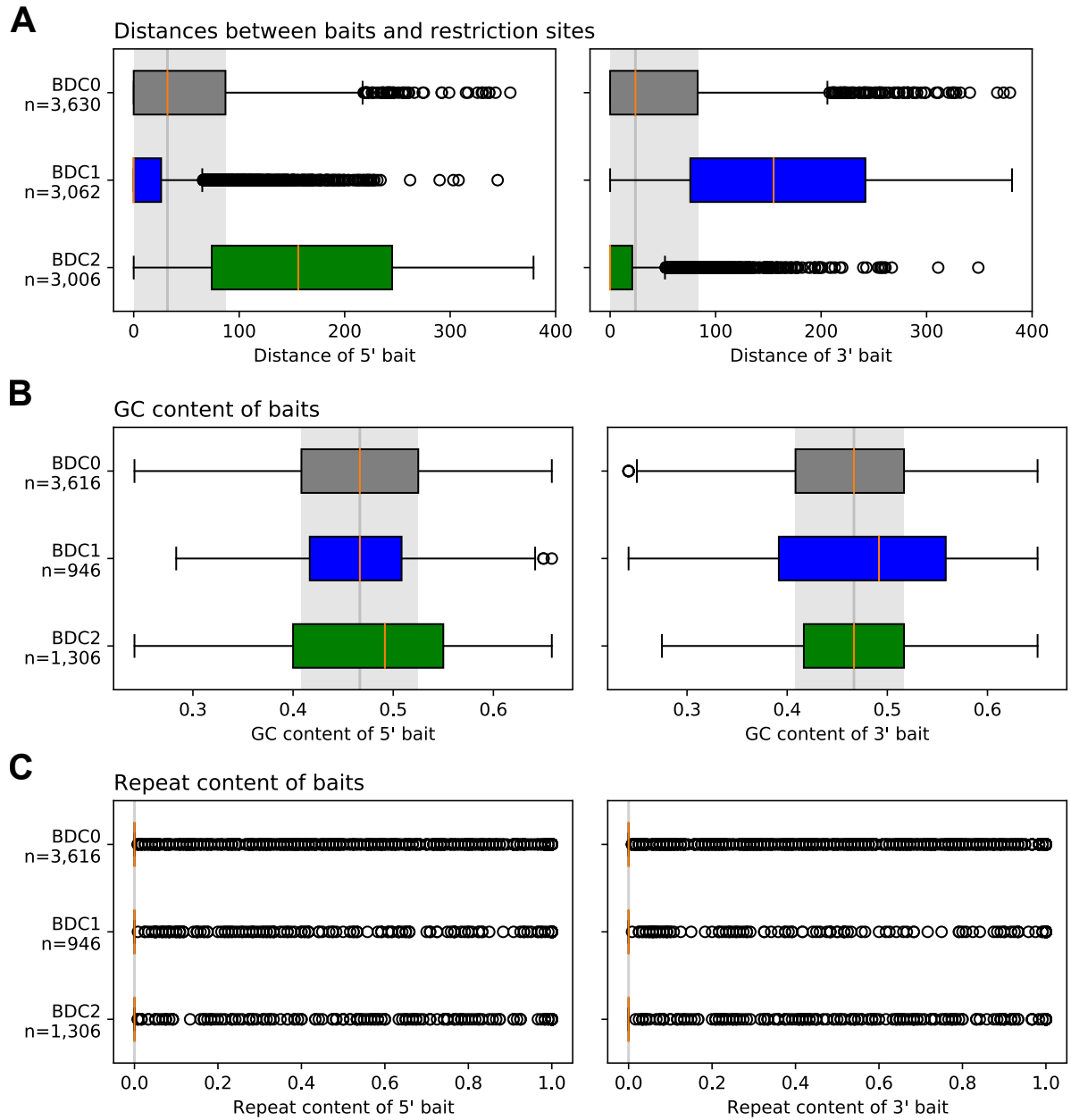
Enrichment of interaction other-ends for ENCODE's Candidate Cis-Regulatory Elements promoters (prom, red), proximal enhancers (enhP, orange), distal enhancers (enhD, yellow), DNase-H3K4me3 (K4m3, magenta) and CTCF sites (blue), as well as for enhancers from the Enhancer Atlas 2.0 (ENA2, purple). REE scores were calculated for interactions identified for the MAC-M0 cell type based on total counts only (light red), total and maximum counts (beige), and maximum counts only (green) using CHiCAGO. In addition, reference REE scores were calculated after randomization of other-ends (gray). This figure shows the results of the same analysis that was performed for the MAC-M0 cell type to obtain Figure 7C in the main text, with the difference that bait-to-bait interactions were discarded. The unnormalized values from the randomization procedure used to calculate the REE scores are shown in Supplementary Table S10.



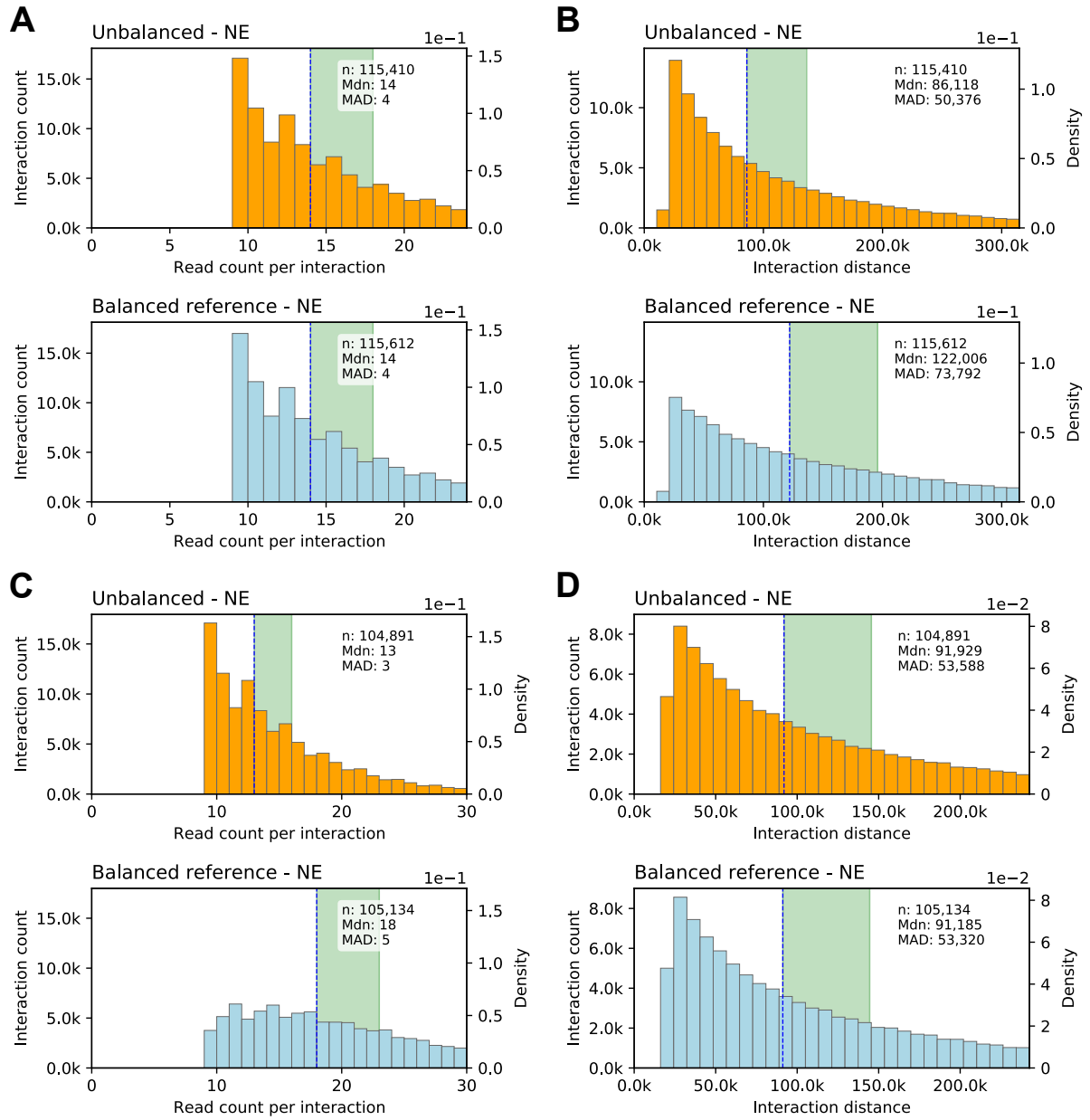
**Figure S12. Unbalanced interactions and their configurations - Mouse embryonic stem cells (mESCs).** This figure shows plots analogous to those shown in Figure 2 in the main text, but these plots are based on data from mESCs (Schoenfelder et al. 2015).



**Figure S13. Overlaps of BFC0, BFC1, and BFC2 fragments - mESCs.** This heatmap is analogous to that shown in Figure S4, but these plots are based on data from experiments with mouse fetal liver cells (MFLC) and mouse embryonic stem cells (MESC). In the one experiments with MESC, RING1A and RING1A/RING1B double-knockout cells were used (MESC-R1ABKO). We determined the proportions of pairwise overlap of fragment sets for all cell types and baited fragment classes. A box in row  $i$  and column  $j$  indicates the proportion of fragments in set  $i$  that are also in set  $j$ .

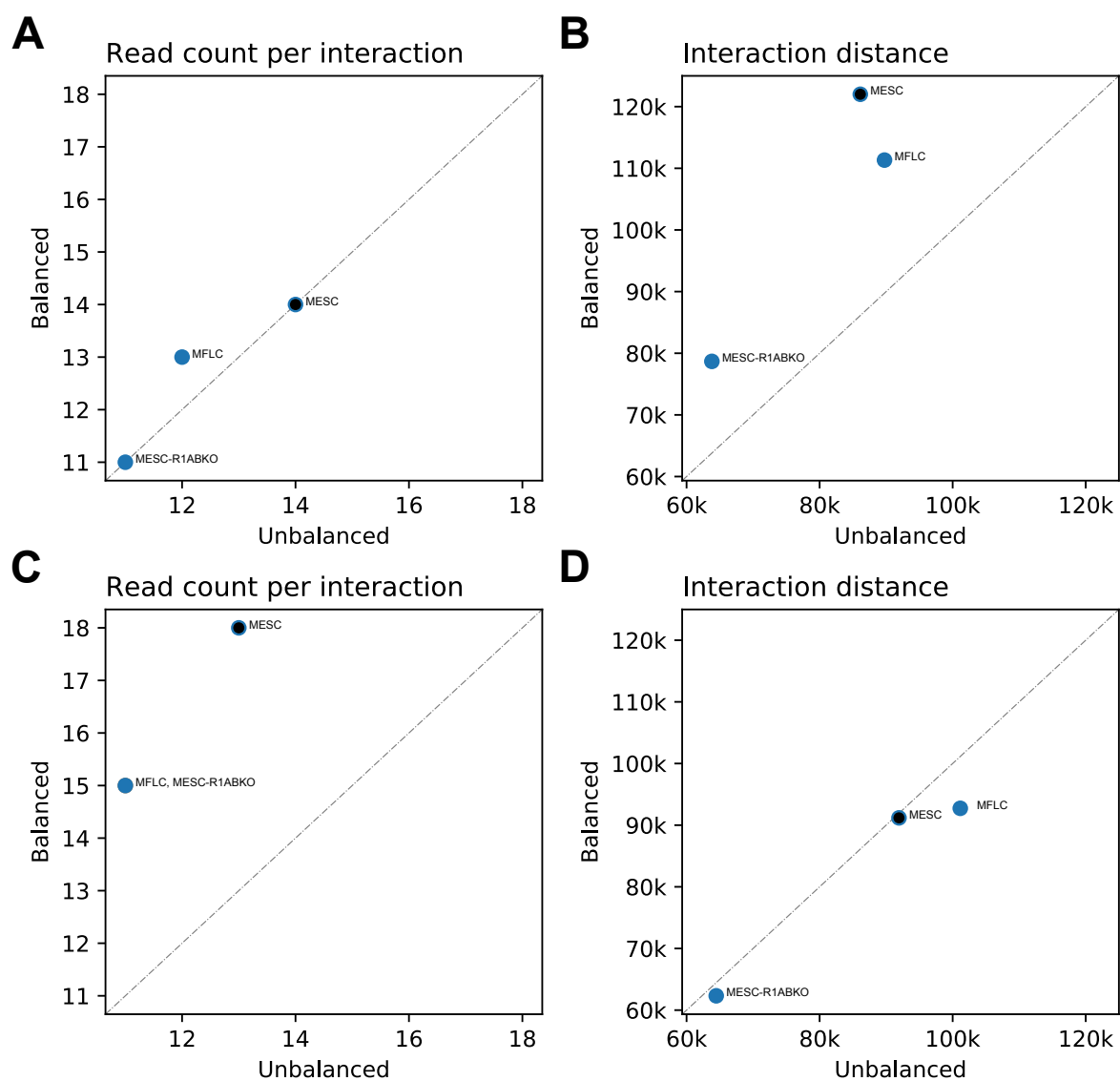


**Figure S14. Bait analysis - mESCs.** This figure shows plots analogous to those shown in Figure 3 in the main text, but these plots are based on data from mESCs. **(A)** Distances between baits and restriction sites. We performed two-sided Wilcoxon signed-rank tests for the classes BFC0 (Rank sum: 3,109,793;  $p = 5.15 \times 10^{-3}$ ), BFC1 (Rank sum: 226,371;  $p \sim 0$ ), and BFC2 (Rank sum: 195,301;  $p \sim 0$ ). **(B)** GC content of baits for the classes BFC0 (Rank sum: 2,829,651;  $p = 2.54 \cdot 10^{-3}$ ), BFC1 (Rank sum: 187,432;  $p = 1.21 \times 10^{-3}$ ), and BFC2 (Rank sum: 349,941;  $p = 4.12 \times 10^{-5}$ ). **(C)** Repeat content of baits. More than 80% of the baits have zero repeat content.



**Figure S15. Impact of technical biases reflected in count imbalances - mESCs.** This figure shows plots analogous to those shown in Figure 5 in the main text, but these plots are based on CHi-C data derived from mESCs.





**Figure S16. Median total read pair counts and distances of interactions - mESCs.** This figure shows plots analogous to those shown in Figure 6 in the main text, but these plots are based on CHi-C experiments with mouse fetal liver cells (MFLC) and mouse embryonic stem cells (MESC). In the one experiment with MESC, RING1A and RING1A/RING1B double-knockout cells were used (MESC-R1ABKO).